DOCUMENT RESUME

ED 283 868 TM 870 396

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE The Golden Rule Agreement is Psychometrically Defensible.
PUB DATE 87
NOTE 19p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Content Validity; Difficulty Level; Item Analysis; *Licensing Examinations (Professions); Predictive Validity; Racial Differences; *Test Bias; *Test Construction; Test Items; Test Reliability; *Test Theory
IDENTIFIERS *Golden Rule (Testing)

ABSTRACT

The agreement between the Educational Testing Service (ETS) and the Golden Rule Insurance Company of Illinois is interpreted as setting the general principles on which items must be selected to be included in a licensure test. These principles put a limit to the difficulty level of any item, and they also limit the size of the difference in difficulty between two groups for any item. Both of these principles are based on sound psychometric theory and practices. The application of these principles may reduce the true differences in ability in order to eliminate bias; in the present state of psychological measurement it is not possible to eliminate all bias in the test and detect all true differences in ability between two groups if the test is biased. The internal consistency of the test is positively related to the amount of bias in the test. Predictive validity, the most important consideration in any licensure test, increases with the elimination of items with aberrant differences in difficulty between groups. Content validity is more than just content relatedness and it is not affected to a greater extent by the agreement than by any other method of item selection. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

4-16-89

FD283868

THE GOLDEN RULE AGREEMENT

IS PSCYHOMETRICALLY DEFENSIBLE

Eulogio Gonzalez-Tamayo

1987

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Eulogio Gonzalez Tamayo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870396

Running Head: GOLDEN RULE

BEST COPY AVAILABLE

SUMMARY

The agreement between the ETS and the Golden Rule Insurance Company of Illinois is interpreted as setting the general principles on which items must be selected to be included in a licensure test. These principles put a limit to the difficulty level of any item adn they also limit the size of the difference in difficulty between two groups for any item. Both of these principles are based on sound psychometric theory and practices. The application of these principles may reduce the true differences in ability in order to eliminate bias; in the present state of psychological measurement it is not possible to eliminate all bias in the test and detect all true differences in abioity between two groups if the test is biased. The internal consistency of the test is positively related to the amount of bias in the test. Predictive validity, the most important consideration in any licensure test, increases with the elimination of items with aberrant differences in difficulty between groups. Content validity is more than just content relatedness and it is not affected to a greater extent by the agreement than by any other method od item selection.

The so-calles "Golden Rule" agreement between the ETS and the Golden Rule Insurance Company of Illinois has been lately in the public forum, again. The main points of the agreement included the identification of a pool of items (1) which at least 40% of blacks and whites answered correctly and (2) for which the rates of right answers of blacks and whites differ by not more than 15 percent. These items should be used exclusively as long as they are available in sufficient number.

A literal interpretation of the settlement and its application to any other case would make not only bad psychometrics but it also would be bad policy. A literal interpretation of the agreement would require (1) to remove from a test all items where the difficulties for two groups differ by 15 percent or more regardless of how apart the two groups are in the attribute measured by the test, (2) if the differences in difficulty are less than 15 percent, the items should be used in the test even if the groups are identical in ability or if the 15 percent favors the low scoring, and conceivably the low ability, group, and (3) if less than 40 percent of one group answered correctly the item, it must be eliminated from the test even if the purpose of the test is to discriminate among the upper 10 percent of the group. It is not possible to defend technically that (1) an item where two groups identical in ability scored 14 percentage points apart in the correct answer is not biased, (2) an item where the low ability group scores less than 15 percent higher than the high ability group is not biased, (3) any difference of 15 percent or more indicates bias regardless of how apart are the two groups in actual ability level, and (4) a test, where all the items are answered correctly by at least 40 percent of the group is useful in discriminating among the applicants in the upper part of the distribution of the ability.

The Golden Rule agreement , however, can also be interpreted as a court case setting the principles by which items should be selected to be included in a

licensure test but without forcing the particulars of the case on every single test and on every single testing situation. With these limitations in mind, the Golden Rule agreement can be defended as based on good psychometric principles.

Principles of the Golden Rule Agreement

The principles included in the agreement can be identified as (1) the items included in a test used for a licensure examination should not exceed a certain level of difficulty to be determined in each testing situation and (2) an aberrant difference in difficulty, a certain number of percentage points in the correct answer, for two groups is an acceptable method of removing items as biased regardless of any other pscyhometric consideration. These two practical principles are based on sound psychometric theory and practice.

Difficulty of the Items

It is a well established principle in psychometrics that items discriminate best, every other thing being equal, when their difficulty levels are at the point of the ability level where differentiation between test takers is wanted. If differentiation is sought at every level of ability, the items must cover the entire range of the difficulty level. If the differentiation is sought at a certain level of ability but not above or below that level, then the items must represent a difficulty level corresponding to that particular level of ability.

In licensure and job selection testing, the most important decision for which the test results are used is to separate those applicants who deserve to be licensed or selected for the job from those who do not. The ranking of the selected applicants is a secondary and less important decision than the separation between qualified and unqualified applicants. Very difficult items discriminate well only

at high levels of ability and easy items discriminate better at low levels of ability. Without a limit to the difficulty level of an item, it is possible to develop tests were none of the applicants will pass the test or all of the applicants will reach a minimum score. These results would not only be embarrassing but also would indicate that the test is useless. The most commnly accepted technique of setting the pass marks in a licensure and job selection test is through absolute standards based on a criterion unrelated to the group being tested. These absolute standards, however, are not completely objective because they are based on personal judgments of experts and these judgments are always somehow subjective.

Standards are obviously wrong when many individuals believe to be competent fail the test or many individuals believe to be incompetent pass the test. By the fallible nature of psychological measurement procedures, a small number of failures among competent people is expected; it is also expected a small number of successes from applicants believe to be incompetent. The standards, however, would be seriously questioned when many individuals believe to be competent fail the test or many individuals believe to be incompetent pass the test. It is impossible to know in advance the percentage of competent applicants taking a particular test; there are, however, some reasonable expectations based on passed experience. A licensure examination has been given for many years with a rate of failure of 30% and the great majority of those who were licensed became competent professionals. If the new standards fail a substantially higher number of individuals among the new applicants believe to be similar to the previous groups, evidently that something may be wrong with the new standards.

Aberrant departures in an empirical case from reasonable expectations are clear indications of errors in the test regardless of how carefully it was developed. Very difficult and very easy items could make the test unreasonably difficult or easy. The best level of item difficulty is a percent of correct answers similar to the percent of applicants expected to pass the test. A 40% passing the item is a

reasonable level of difficulty for the two licensure tests included in the litigation. A preliminary study showed that the pass rates for whites were 78 percent in the life insurance license and 82 percent in the accident and health insurance license; the percentages for blacks were 65 and 55 respectively. The 40 percent passing the item from the agreement, makes the criterion of item selection somewhat harder than the ideal for the lower ability group (55 percent passing for blacks in the accident and insurance license).

The principle of the Golden Rule agreement can also be defended without any consideration of the pass marks. A 40 percent passing the item indicates a difficulty slightly above the middle difficulty level; these kinds of items provide better differentiations among the apllicants than more difficult items. In a multiple choice test with four alternatives for each item, it is expected that, as an average, 25 percent of the applicants from any group will answer correctly any item by merely guessing even if none of the members of the group knows the correct answer. As a rule of thumb, any item of the four choices type answered correctly by less than a third of the persons in the group would not measure any ability but merely error. The 40 percent difficulty level of the Golden Rule agreement seems to be a principle based on sound psychometric theory.

The application of this Golden Rule principle will not lead by itself to control from the outset of what the test results will be as a superficial reading of the agreement might indicate. The 40 percent right answers does not force the test developer to license or select any number of applicants since the passing score is based on the difficulty level for the minimally competent applicant rather than on the percentage of applicants passing the item. The principle does not establish either what the average score will be for any of the two groups; the principle sets what the results are not supposed to be rather than what they are going to be. The items should not be so difficult

that the average score for any group is below the 40 percent of correct answers and this is a sound psychometric principle since a lower mean score would include too large an error.

Aberrant Differences Passing the Item

The Golden Rule agreement also establishes the principle that, for the items selected to be included in the test, the rate of right answers of blacks and whites may not differ by more than 15 percent. The average difference between the two groups in passing the test was 13 percent for the life insurance test and 27 percent for the accident and health insurance license test.

The difference in right answer rates for an item to be removed from the test should not be a fixed number applicable to any test or to any testing situation; rather, it should be based on the test characteristics and the groups under consideration. The 15 percent of the agreement can be interpreted as representing an aberrant result in difficulty level differneces for the two groups of the particular case addressed by the agreement; in another case the percentage could be higher or lower. The agreement establishes the principle that differences in passing rates is a legitimate criterion to consider an item as biased and to remove it from the test. This principle makes many psychometricians very unconfortable since the p method technique of detecting bias does not have many supporters.

An outstanding issue of item bias techniques is the effect of removing the biased items on test characteristics; specifically, whether or not the removal reduces the adverse impact of the test on some groups without impairing its predictive validity (Iroson, 1982). Unfortunately, statistical procedures for detecting biased items have become an end in themselves where the results are carefully tabulated and reported and nothing further is done (Scheuneman, 1982).

The controversy about test bias started, to some extent, because large group differences in test scores were found between blacks and whites. If the removal of biased items does not reduce the mean score difference between the groups, there seems to be no compelling reason to use any psychometric technique of detecting items as biased.

The single most important consideration in licensure and selection tests is their predictive validity. Present methods of determining the validity of the test through content validation are inadequate; they are safe and cheap to implement but, unfortunately they are design to confirm. Test constructors avoid relating knowledge and skills measured to actual performance saying that these kind of studies are impractical because the test goes into operation before the validation can take place; or is it because people are afraid of the consequences of what they may find or, more accurately, perhaps, not find (Madaus, 1986)?

Nungester (1977) reported that removal of items with unacceptable discrimination or difficulty indices, difficulty measured as simple percentage passing the item, improved the reliability and validity of the tests to a greater extent than the removal of items detected as biased by three psychometric methods where the change in reliability and validity was insignificant. Gonzalez-Tamayo (1984) reported that (1) it is possible to reduce the mean group difference by eliminating item detected as biased by the simple method of percentage passing the item without lowering the validity of the test; the removal of biased items produced both a subtantial reduction in mean differences between the majority and the minority groups and an increase in the predictive validity of the test, and (2) a simple criterion of mean group difference in difficulty plus one half of one standard deviation to determine an item as biased worked as well as an a priori criterion in detecting biased items.

The acceptance or rejection of a particular psychometric technique is based, very often, on the theoretical sophistication involved rather than in the

empirical effects of its use. Methodology can be elevated from handmaiden to the master of the process of inquiry (Nisbert, 1976). Some years ago, Lawshe (1969) found that technically more elegant methods were not better than simpler ones in estimating the reliability, the internal consistency, and the weighting combination of scores and ended the analysis with a warning:" the statistical nicety of the present methods suggests or implies an order of precision which is not basically inherent in the data; psychological measurements, at this point in time, are quite unreliable; to suggest otherwise by using unwarranted degrees of statistical precision is for the psychologist to delude others and, perhaps, to delude himself (p.122)".

Misconceptions of the Golden Rule Principles

Some implications of the principles established by the Golden Rule agreement have been presented in sicentific publications that may be considered as misconceptions or misinterpretations that do not stand under severe scrutiny of the rationale on which the principles are based.

Actual Group Differences and the Golden Rule Agreement

The elimination or reduction of differences in average group mean test score is a direct result of the application of the principle of removing items where the right answers rate difference between the groups is more then 15 percent. Selecting items where the passing rates between blacks and whites differ by 15 percent or less, or rejecting items where the passing rates difference is larger than 15 percent, will reduce the actual differences shown in the test. This fact may conceal the true differences in ability between the two groups and psychometricians may be more amenable to other techniques of detecting bias where

the ability level of the groups is supposedly controlled. The problem is that nobody knows the actual size of the true differences in ability between two groups regardless of how these differences were developed. A person's score on the total test is directly related to his or her score on each item. If there are several biased items, that will pull the average minority person's score down so that the test does not really compare "like" groups and the methods will underestimate the test bias. If we do not know the actual differences in ability between two groups we can not use a biased test to classify them; if we know the actual differences in ability we do not need a test to measure them.

The choices for eliminating bias in the test are not many, nor are they clear or pleasant. Some persons may prefer to control for ability at the risk of underestimating bias; in such a case, the items removed from the test will be those which showed differences caused by irrelevant characteristics raher than by real differences between the groups. Other person may prefer to eliminate all bias at the risk of possibly eliminating also some real differences and in that case they will use the principle sponsored by the Golden Rule agreement of a certain percentage difference in passing rates. The choice can not be based on psychometric principles, none of the alternatives is completely acceptable from a technical point of view, but rather it is a question of values or beliefs. In the present development of our craft, we can not do both things at the same time, eliminating all bias in the test and detecting all true differences in ability between the groups.

Reliability of the Test and the Golden Rule Agreement

The Golden Rule agreement does not allow to use any item where the passing rates difference between two groups exceeds a certain predetermined percent. The rule, in that sense, seems to reduce the possible range of scores and, as a

result, it may also lower the reliability of the test. Every other thing being equal, the reliability of a test is positively related to the variability of scores. The lowering of the reliability by the application of the Golden Rule principle is, however, only a possibility rather than an empirical fact. This possibility represents a serious consequence that deserves a critical analysis.

Imagine two groups of identical ability at every point of the ability continuum and not only as an average; the case represents a perfect match of every single individual from one group with an individual in the other group. This is an unrealistic case but useful in analyzing the effects of biased items. Table 1

Table 1 about here

---

reports the theoretical data needed for an analysis of variance of fourr different cases that will be studied during the discussion.

The analysis of variance is selected to calculate the intraclass correlation coefficient as an indicator of the reliability of the test. The procedure allows to calculate two reliability coefficients depending on the kind of decisions for which the test results will be used. If the test scores are used to rank the applicants, norm-referneced decisions, the reliability coefficient will be identical to the KR_{20} for items scored dichotomously. If the test results are used for criterion-related decisions, to select or reject, to license or to deny the license, then the reliability coefficient calculated through the procedures developed by the generalizability theory will be lower than the KR_{20}. The difference between the two intraclass correlation coefficients is given by the variance attributable to the items effect. In norm-referneced decisions, item variability will not affect the decisions because all the applicants are tested with the same test and any systematic item variability will affect equally to all the applicants.

item variability in norm-referenced decisions is not error. In criterion-referenced decisions, item variability is error because reaching or failing to reach a predetermined criterion may be the result of the difficulty level of few items; in any other situation, different items selected under the same procedure may produce different scores and different results in relation to the criterion and item variability is, therefore, an error affecting the decisions.

The different sources of variability for the analysis of variance are reported in table 2 where the R_1 indicates the reliability coefficient for norm-referenced

Table 2 about here

decisions and R_2 indicates the reliability for criterion-referenced decisions. The data in table 2 from the data in table 1 were obtained through the standard procedure for a two-way analysis of variance where persons and items are the two dimensions. The intraclass correlation coefficients were calculated through the formulas developed in the generalizability theory.

Case I represent an analysis of variance for one single sample composed of the individuals from group I and from group II. The total $\sum x^2_{.i}$ was calculated from the total number of persons answering correctly each item rather than adding the totals from the two groups.

Case II contains 10 additional items considered as compeltely biased because every single person of group I answered correctly every single item and none of the applicants from group II answered correctly any of the 10 items.

Case III is based on the assumption that the original test of 30 items was completely biased against group II because none of the applicants from this group answered correctly any of the 30 items. The test discriminates among the applicants

of group I but it does not discriminate among applicants of group II.

Case IV represents a more extreme case where all the items of the original test are answered correctly by all the applicants of group I and none of the items is answered correctly by any of the applicants of group II.

The data of table 2 clearly indicate that the greater the amount of bias in the test against one of the groups the higher the reliability of the test. If two groups are identical in ability (1) a test that faithfully detects this identity in ability, case I, will provide an internal reliability coefficient of 0.54; (2) a test where only part of the items are biased, case II, will provide an internal reliability coefficient of 0.78; (3) a test where all the items are biased against one of the groups, cases III and IV, will provide a reliability coefficient of 0.97 and 1.00 respectively.

The conclusion from this analysis is that reliability and bias are two different concepts but they are positively related: the greater the amount of bias in the test, the higher the reliability even among groups identical in ability. Bias is an error of measurement but it is not a chance error. Bias is part of the true score rather than of the error. The true score of a test includes any error of measurement that is not chance error; by definition of error, any individual idiosincracies not related to the attribute but rather with the form the examinees answer the items and which may affect the observed score, is included in the true score. When bias is introduced in a test, it affects the true score and the reliability coefficient, the ratio of true to observed score, will increase because in any improper fraction when both the numerator and denominator are increased by the same amount the entire fraction increases. The greater the amount of bias the greater the heterogeneity of scores because the bias increases the true variance; the greater the heterogeneity of scores, regardles of how omegeneous or heterogeneous the groups are, the higher the reliability.

Psychometric techniques were developed to detect individual differences and the most reliable psychometric instrument is one that produces the largest individual differences among the examinees. The concept of bias is predicated on the premise that the test may not measure equally well two groups with different experiences. All the statistics from the classical test theory are gorup dependent. Test are good at ranking individuals within a very homegeneous group but they are not good at ranking heterogeneous groups (Goodenough and Harris, 1952). It may be that statistics are not good to study heterogeneous groups at all.

Test Validity and the Golden Rule

Linn and Drasgow (1987) reported that the Golden Rule agreement, if applied, "will degrade validity, because it is precisely those items that provide the best measurement of the underlying attribute being measured that are most likely to fail to meet the criteria (p.8)". This may or may not be the case depending on the circunstances. No empirical data are presented in support of their conclusion; empirical data, however, rather than theoretical speculation, should provide the answer to the problem.

In terms of predictive validity, as reported earlier (Nungester, 1977; Gonzalez-Tamayo, 1984), the removal of items with unacceptable difficulty indices differences among the groups improved the validity of the test. Those claiming that the application of the Golden Rule will most likely reduce the validity of the test, have the responsibility of providing the supporting empirical data; even thoug empirical data are available, nobody has reported them.

In terms of content validity, it may be said that, perhaps, it is not possible to construct a test that still has content validity, that is, the content of the items adequately reflects the knowledge base that is being measured, if items are discarded solely on the basis of differences in the answers of blacks and

whites. This is a possibility that may occur any time that items are eliminated by any procedure of item selection and the Golden Rule principle is only an additional procedure of item selection. It is unconceivable that test developers can not produce an adequate number of items in an area, domain, or subdomain, where the difficulties for two groups are not aberrant or extreme. Finally, the conept of ontent validity may be misplaced since most of the tests provide items that are content related. Content validity is more than content relateness and it requires that the exercises included in the test must represent or reproduce the essential characteristics of the universe in their proper proportion and balance (Cronbach, 1971). Present methods of determining content validity are safe and cheap to implement but they are inadequate. The concept of sampling of items from a universe has been criticized as impractical (Thorndike, 1966) and even improper (Loevinger, 1965). In lincesure examinations, content validity is the only validity available; eliminating the concept of sampling or leaving it to the test developer, would reduce the test to a series of exercises that are job related but with very limited meaning or usefulness.

BIBLIOGRAPHY

Cronbach, L.J.,1971. Test Validation. In R.L. Thorndike (Ed.) Educational Measurement, 2nd Edition. Washington, D.C.: American Council on Education.

Gonzalez-Tamayo, E.,1984. An Item Content Criterion of Debiasing a Test. ERIC Documents Reproduction Service. No.Ed. 247-240.

Goodenough, F.L. & Harris, D.B., 1950. Studies in the Psychology of Children's Drawings:II, 1928-1949. Psychological Bulletin, 47, 369-433.

Ironson, G.H.,1982. Use of Chi-Square and Latent Trait Approaches for Detecting Item Bias. In R.A. Berck (Ed.) Handbook of Methods for Detecting Test Bias. The Johns Hopking University Press, Baltimore, Maryland.

Lawshe, C.H., 1969. Statistical Theory and Practice in Applied Psychology. Journal of Personnel Psychology, 22, 117-124.

Linn, R.L. & Drasgow, F., 1987. Implications of the Golden Rule Agreement. The Score. Volume IX, No.4.

Loevinger, J., 1965. Person and Population as Psychometric Concepts. Psychological Review, 72, 143-155.

Nisbert, R., 1976. Sociology as and Art Form. New York: Oxford University Press.

Nungester, R.J., 1977. An Empirical Examination of Three Models of Item Bias. Doctoral Dissertation, Florida State University.

Madaus, G.F., 1986. Measurement Specialists: Testing the Faith. A Reply to Mehrens. Educational Measurement Issues and Practice, 5, 11-14.

Scheuneman, J.D.,1982. A Posteriori Analysis of Biased Items. In R.A. Berck (Ed.) Handbook of Methods for Detecting Test Bias. The John Hopkings University Press, Baltimore, Maryland.

Thorndike, R. L., 1966. Reliability. In A. Anastasi (Ed.) Testing Problems in Perspective. Washington D.C.: American Council on Education.

Table 1

Total Scores, Sums of Squares of Individual Scores and of Item Responses and Correct Responses to Each Item for Two Identical Groups

Source	Group I	Group II	Total
N	100	100	200
ΣX	1,900	1,900	3.800
$\Sigma X^2_{p.}$	37,532	37,532	75,064
$\Sigma X^2_{.i}$	121,500	121,500	486,000(1)
Items	**Correct Responses**		
10	70	70	140
10	65	65	130
10	55	55	110

(1)This total is not the sum of the squares for each group but rather the sum of squares of the responses to each item.

Table 2

Analysis of Variance Data for Four Cases Based on the Information of Table 1

	Source of Variance	Variance Data			Intraclass Correlation	
		Sum of Squares	d.f.	Mean Squares	R_1	R_2
I	Total	1393.3333	5999			
	Persons	95.4667	199	0.4797		
	Items	23.3333	29	0.8045		
	Pxl	1274.2000	5771	0.2208	0.5397	0.5364
II	Total	1920.0000	7999			
	Persons	196.6000	199	0.9879		
	Items	50.0000	39	1.2821		
	Pxl	1673.4000	7761	0.2156	0.7818	0.7776
III	Total	1298.3333	5999			
	Persons	649.3999	199	3.2633		
	Items	5.8333	29	0.1496		
	Pxl	643.1001	5771	0.1114	0.9659	0.9659
IV	Total	1500.0000	5999			
	Persons	1500.0000	199	7.5377		
	Items	0.0000	29	0.0000		
	Pxl	0.0000	5771	0.0000	1.0000	1.0000

Case I: The two groups scored identical on the test

Case II: Same as case I but 10 completely biased items were added

Case III: All applicants in group II scored zero.

Case IV: All applicants in group I scored perfect and all in group II scored zero